# Compilation and Annotation of the Discourse-structured Blog Corpus for German

**Holger Grumt Suárez, Natali Karlova-Bourbonus, Henning Lobin**
Department for German Linguistics and Literature
Applied and Computational Linguistics
Justus-Liebig-University Giessen, Germany
Holger.H.Grumt-Suarez@germanistik.uni-giessen.de, natali.karlova-bourbonus@zmi.uni-giessen.de,
henning.lobin@germanistik.uni-giessen.de

**Abstract**

The present paper reports the first results of the compilation and annotation of a blog corpus for German. The main aim of the project is the representation of the blog discourse structure and relations between its elements (blog posts, comments) and participants (bloggers, commentators). The data included in the corpus were manually collected from the scientific blog portal SciLogs. The feature catalogue for the corpus annotation includes three types of information which is directly or indirectly provided in the blog or can be construed by means of statistical analysis or computational tools. At this point, only directly available information (e.g., title of the blog post, name of the blogger etc.) has been annotated. We believe, our blog corpus can be of interest for the general study of blog structure or related research questions as well as for the development of NLP methods and techniques (e.g. for authorship detection).

**Keywords:** CMC, blog corpus, corpus compilation, corpus annotation, TEI

## 1. Introduction

In our opinion, two views on computer-mediated communication (CMC) – linguistic and structural – have so far been established. According to the linguistic view, the language of CMC represents a distinct type of language form besides written and spoken language. Moreover, it combines characteristics of these two traditional language forms thus constituting a bridge between them. The structural view in its turn concentrates on building up of CMC. Two different kinds of CMC structure can be distinguished – external and internal. External structure relates to the representation, or layout, of CMC by means of HTML mark-up language which may be an individual decision of a developer. External structure most of the blogs includes for example a header (title), content, a footer (contact information) and a sidebar (site navigation). Internal structure in its turn relates to the generic structure of the CMC content. It describes a set of structural elements (e.g., post, comment, thread, word cloud etc.), properties and principles a CMC is constructed of and built on to function as a holistic construct and to match its purpose.

The identification of the full spectrum of CMC characteristics – linguistic or structural – still faces some major challenges primarily as a result of lacking valid annotated data. Storrer (2014: 189) claims that for this purpose a special – third - kind of corpus besides the written and spoken corpora is needed. She also adds that appropriate standards, methods and quality criteria for the study of CMC are crucially important as well.

In the present study, the structural nature of the weblog (henceforth blog) as a representative genre of CMC is of interest. We describe the genre blog as a dynamic, "living" construct of interrelated and interacting elements. The dynamics of a blog arise from its constant expansion as a result of ever more comments and blog posts as well as on the account of new blog participants. Additionally, the author of the blog (henceforth blogger) can edit his post any time and add new information on request. The interrelatedness and interaction between elements (blog post, comments) and agents (blogger, commentators) of the blog contribute to the dynamics of the blog as well.

To demonstrate this idea, we compiled the first version of an annotated blog corpus in German using the scientific blog portal SciLogs (SciLogs, 2016) as a data source. The corpus includes both blog posts and related comments. The catalogue of features for the annotation of the corpus is based on three types of information directly or indirectly available from the data source. The typology of information is proposed in Section 3.2.1.

The structure of the paper is as follows. Section 2 provides an overview of the studies related to the topic of the present project. Section 3 describes the main steps conducted for the purpose of the blog corpus compilation and annotation. Some observed challenges for the automation of the task and possible solutions are also included in this section. Finally, Section 4 reports the results of the project and outlines the next steps.

## 2. Related Work

Currently, there is a limited number of publicly-available, large-scale blog corpora. This is surprising given the great influence of blogs on the web in general.

An example for one of the few large-scale blog corpora is the Birmingham Blog Corpus compiled at Birmingham City University. The corpus consists of more than 630 million words, including a 180 million words sub-section separated into posts and comments (Kehoe, 2012). One objective of this corpus was to analyze if "comments could be used to improve document indexing on the web" (ibid.). The online tool (WebCorp, 2016) of the Birmingham blog corpus allows the querying of words and phrases, but there is no possibility to either search the comment structure, a specific time period, keywords or a specific blogger respectively commentator.

Another example of a blog corpus is the bilingual (German, French) corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne), which is part of the LETEC (Learning and Teaching Corpus) (Abendroth-Timmer, 2014). This corpus is included in the CoMeRe (Communication médiée par les réseaux) project, which "aims to build a Kernel corpus assembling existing corpora of different CMC […] genres and new corpora build on data extracted from the Internet" (Abendroth-Timmer, 2014). The INFRAL blog corpus consists of posts from two groups: a group of ten francophone learners of German as a foreign language from l'Université de Franche-Comté and a group of nine German-speaking learners of French as a foreign language from the University of Bremen who e.g. had to discuss various intercultural topics. One task of this corpus was the modeling of the structure of interactions. Therefore, every comment has been given a reference to the ID of the post, but the links between the comments themselves are not included. The TEI schema developed for the CoMeRe project – this project is also part of the TEI special interest group (SIG) "computer-mediated communication" (CMC) – will be an important basis for our own schema.

Finally, the German language wordpress blog corpus by Barbaresi and Würzner (Barbaresi & Würzner, 2014) is another example of a blog corpus worth mentioning. The corpus consists of a total of 158,719 German wordpress blogs. The collected data is released under the Creative Commons license. The corpus can be used for example in the lexicography for the purpose of dictionary building. Moreover, it can be a good source "to test linguistic annotation chains for robustness" (Barbaresi & Würzner, 2014).

# 3. Methodology

## 3.1 Data Collection

To date, we have compiled a test corpus in the German language, which contains 21 blog posts and 195 comments related to those blog posts. For this test corpus, we wanted to cover a whole week and we therefore randomly choose week 49 in 2015 (from November 30, 2015 to December 6, 2015). The source of the data is the scientific blog portal SciLogs (SciLogs, 2016). SciLogs is subdivided into different sections (BrainLogs, ChonoLogs, KosmoLogs, WissenLogs) where scientists – and those interested in science – can interact in interdisciplinary discussions about science. For the test corpus, we did not focus on a particular section; we extracted the blogs from different sections.

The result of analyzing the SciLogs source code is that the different sections store the data using the same template. In the source code, we can find the information for title, category, keywords, date, name of the blogger / commentators, the comments and their different levels of indentation, the permalinks of the comments and the blogpost itself. The data collection for the test corpus was done manually and in the process of the work we discovered some complications that will have to be dealt with later during the automation phase. We will discuss some of these complications in Section 3.2.3.

Our next step will be to complete our corpus with the data appeared in 2015 considering all SciLogs sections. According to our current knowledge, the SciLogs data of 2015 includes about 1.200 blog posts and 12.000 comments. Retrieval of the blog data from the web will be conducted semi-automatically. For this purpose, an open source program HTTrack Website Copier (Roche, 2016) will be used. HTTrack enables the download of all kinds of the website data stored on the server including HTML pages, images and other files to a local directory on a computer. After the retrieval step, the data will be cleaned from the noise in the data and represented in form of HTMl pages (external structure). Finally, the relevant content will be extracted from the HTML pages and annotated with TEI annotation standard (internal structure). The programming language Python and its packages for XML parsing will be used for this purpose.

## 3.2 Data Annotation

### 3.2.1 Types of Blog Information

We distinguish between three types of information provided in the blog based on how the former is made available. The first type (**type A**) incorporates information which is directly available in the blog or from the source code of the blog site. In the blog post structure, it includes the blog post itself along with the meta information such as the title of the blog post, date of creation, the name of the blogger, the categories the entry belongs to and main keywords. In the structure of the comments, type A information is represented by the total number of comments as well as the name of the commentator, date and comment ID. The second type (**type B**) includes information which is not directly available but can be inferred from **type A** information, e.g. usual activity time of a commentator (at what time a particular commentator usually writes his comments). Finally, the third type (**type C**) is an interpretative information type. This kind of information is neither directly nor indirectly provided in the blog but is rather the result of statistical (basic statistics), linguistic (e.g., part-of-speeches) and discourse (e.g., topic identification with topic modeling) interpretation and analysis of the blog entries. The interpretative information type can either be collected manually or by use of computational tools.

### 3.2.2 Annotation Standard

To date, no standard exists for representing CMC data. One option could be to design an XML schema for CMC from scratch, which would perfectly fit the needs of our project. The main reason as to why we are not going along with XML is that the schema would be idiosyncratic and the corpus would not interoperate without causing difficulty with other resources. When searching for a standard for the representation of texts in digital form, one

will take a look at the Text Encoding Initiative (TEI). However, none of the modules in the current version of the TEI Guidelines (P5) can be adopted for our project. Fortunately, the SIG CMC group under the direction of Beißwenger (TU Dortmund) has been working on the adaption of TEI guidelines to the presentation of genres of CMC since 2012 (Beißwenger, 2015). Given that no module for CMC is so far ready to use, we have started to look for schema drafts by the SIG CMC group and up to now, a couple of corpora have been released by the SIG CMC group. Among them are CMC genres like tweets, email, text chat, wiki discussions and weblogs (Chanier, 2014; Beißwenger, 2013; Storrer, 2015). The schema that fits our needs best, is the one released in 2014 by the French network CoMeRe (Communication médiée par les réseaux) (Hriba, 2013). The CoMeRe schema is based on the previous schema draft by DeRiK (Beißwenger, 2013) and includes e.g. the metadata schema for CMC. But still, there is no possibility for representing the full structure of a blog and especially the related comments. Our goal is to take the latest schema draft provided by the SIG CMC (Beißwenger, 2016) and not to try to change the main characteristics of the schema. The status of that schema is that of a "core model for the representation of CMC" (Beißwenger et al. 2012: 6). And so we will need to redefine some elements while also introducing some new ones.

#### 3.2.3 Challenges and Possible Solutions

A number of aspects are challenging since the task of blog corpus annotation is in some cases the result of the particularities of the content management system (CMS) functionality used by our blog data source. Most of the challenges deal with the structure of the comments. As we are at an early stage of our project, only a limited number of challenges and solutions will be described here.

The first challenge is due to the absence of an editing function for the comments. The commentator who edits the text of the comment creates a new entry which appears in the timeline as an autonomous comment. Thus, the comments structure of our blog corpus includes both original comments and their edited versions appeared to the time of the data collection. Though, this aspect does not have an impact on the difficulty of the automation of the annotation task. However, it first impacts the accuracy of the total number of distinct comments (type A information). Second, it creates confusing linkages in the comments structure.

The latter problem also arises as the result of the second challenge – the possibility that one comment refers to more than one previous comment. Unfortunately, the CMS of our blog source does not offer any special options to mark or highlight multiple comment references. In some cases, the commentators use constructions such as [@name]* to overcome this problem. In other cases, an additional analysis of the comment content is required. For the purpose of the study, only explicit references are taken into consideration. No deeper content analysis has been conducted. The identification of multiple references and their annotation with TEI was processed automatically and then manually checked for mistakes in order to achieve accurate and reliable results. We believe that it is be less time- and cost-consuming than fully manual processing of the data. The automatic part is conducted based on explicit marks of multiple reference such as [@name]*. In the TEI blog annotation the multiple references are specified by enumeration of the ids of their comments (<replyTo>).

Finally, the third challenge is the task of the correct assignment of the comments to the level in the hierarchical structure of the comments. At present, the number of possible level assignments is limited to five. All comments appearing after the first comment on the fifth level are (wrongly) assigned to the fifth level. In order to solve this problem, we developed a simple algorithm to compute the correct level of the comments. The algorithm first takes the person reference ("@name", "[name] schrieb (engl.: wrote)" etc.) included in the text of the analyzed comment as the input. In the case of multiple references, only the first reference is taken into consideration. The algorithm then searches backwards for the matches between the person reference and the name of the commentator in the previous comments. Through matches, level of the analyzed comment is computed as the sum of the level assignment of the comment which the person reference belongs to and 1. By absence of the references, the level of the comment is counted subsequently.

## 4. Results

The main steps conducted for the purposes of a scientific blog corpus compilation as well as challenges faced during this process were described in the present study. The current version of the corpus contains 21 blog posts and 195 related comments written in the period of one week. We are convinced that comments are an essential part of a blog corpus. On their own or in connection with the correspondent blog post, they provide valuable information for processing diverse research questions on the language of the blog and its structure. For example, based on the name of the commentator and the time of his comments, we can compute at what time a particular commentator is active in the blog.

The data for our blog corpus was manually collected and annotated according to the TEI schema drafts developed by the TEI special interest group. For the annotation, three types of information (direct, indirect and interpretative) based on the availability of the latter have been identified. The present version of the corpus includes annotation of the first type - directly retrieved information (e.g., the name of the blogger, title of the blog entry, the name of the commentator etc.). The next objective of the project is an expansion and full annotation of the corpus as well as the automation of the data collection and annotation task. At the final stage of the present project, our annotated corpus will be made available to the interested community to perform diverse kinds of research and experiments. Our aim is to enable the access to the corpus through a

searchable online database. Additionally, we plan to make a part of the corpus to be available upon request. For the legal aspects of the SciLogs data usage and publication an external competent institution will be consulted.

## 5. Acknowledgements

We would like to thank our anonymous reviewers for their insightful comments and suggestions. Following the feedback, we included several improvements in our paper.

## 6. References

Abendroth-Timmer, D. et al. (2014). Corpus d'apprentissage INFRAL (Interculturel Franco-Allemand en Ligne). Banque de corpus CoMeRe. *Ortolang.fr: Nancy*. https://hdl.handle.net/11403/comere/cmr-infral (last retrieved 23 August 2016).

Barbaresi, A., Würzner, K.-M. (2014): For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In KONVENS 2014, NLP4CMC workshop proceedings, p. 2–10.

Beißwenger, M. et al. (2012). A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative (jTEI), Issue 3.

Beißwenger, M. et al. (2013). DeRiK: A German reference corpus of computer-mediated communication. pp. 531-537. In: M. A. Finlayson (Eds.), LLC. The Journal of Digital Scholarship in the Humanities, Volume 28, Number 4. Oxford, OUP, pp. 531-537.

Beißwenger, M. (2015). Computer-Mediated Communication SIG. In TEI Website. http://www.tei-c.org/Activities/SIG/CMC/ (last retrieved 20 April 2016).

Beißwenger, M. (2016). SIG:Computer-Mediated Communication. In TEI Website. http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication (last retrieved 20 April 2016).

Chanier,T. et al. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. In: Special issue on Building And Annotating Corpora Of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics. Journal of Language Technology and Computational Linguistics. Berlin, GSCL, pp1-31.

Hriba, L., Chanier, T. (2013). Projet européen TEI-CMC. Comere: Corpuscomere. Communication médiée par les réseaux. In Comere Website. https://corpuscomere.wordpress.com/tei/ (last retrieved 20 April 2016).

Kehoe, A., Gee, M. (2012). Reader comments as an aboutness indicator in online texts: introducing the Birmingham Blog Corpus. In Studies in Variation, Contacts and Change in English 12: Aspects of corpus linguistics: compilation, annotation, analysis. http://www.helsinki.fi/varieng/series/volumes/12/kehoe_gee/ (last retrieved 20 April 2016).

Roche, X. (2016). HTTrack. Website Copier. http://www.httrack.com/ (last retrieved 20 April 2016).

SciLogs (2016). SciLogs. Tagebücher der Wissenschaft. Spektrum der Wissenschaft Verlagsgesellschaft mbH. http://www.scilogs.de/impressum/ (last retrieved 20 April 2016).

Storrer, A. (2014). Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In: A. Plewina & W. Andreas (Eds.), Sprachverfall? Berlin, De Gruyter, pp. 171-196.

Storrer, A. (2015). ChatCorpus2CLARIN: Integration of the Dortmund Chat Corpus into CLARIN-D. In CLARIN-D Website. http://de.clarin.eu/en/curation-project-1-3-german-philology (last retrieved 20 April 2016).

WebCorp (2013). Birmingham Blog Corpus. WebCorp: Linguist's Search Engine. Birmingham City University. http://wse1.webcorp.org.uk/cgi-bin/BLOG/index.cgi (last retrieved 20 April 2016).